September 2024

Building an Evidence-Based Foundation to Understand Foreign-Born Scientists and Engineers' Participation in the US Workforce

Final Report

Project Performer: NORC at the University of Chicago

Period of Performance: February 2022- September 2024

The America's DataHub Consortium (ADC), a public-private partnership is being utilized to implement research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). This report documents research funded through the ADC and is being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed in this report do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. This product has been reviewed for unauthorized disclosure of confidential information under NCSES-DRN25-024

Background and Purpose

This project is part of the National Secure Data Service (NSDS) Demonstration Project conducted by the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation. Authorized under the 2022 CHIPS and Science Act, the NSDS Demonstration aims to inform a government-wide effort to strengthen data linkage and access infrastructure. This initiative aims to facilitate statistical activities that support enhanced evidence-building for the American public.

Data sharing between governmental and non-governmental agencies is a complex process involving vast quantities of data, varying levels of sensitivity, and significant legal implications. Developing a shared service under such conditions requires a system offering high levels of both efficiency and security. A five-part series of awards, collectively titled "Foreign Born Scientists and Engineers in the Workforce (FBSE)," serves the purpose of gaining knowledge about this population while exploring different components of data sharing.

This specific project, titled "Evidence-Based Foundation to Understand Foreign-Born Scientists and Engineers' Participation in the U.S. Workforce," is one part of the series led by NORC at the University of Chicago (NORC). There are limited data sources representing the FBSE, making it a difficult population to measure. To assess how data sources might address key questions related to FBSE's, the project consisted of both a research track and a demonstration track. The research track aimed to fill knowledge gaps surrounding FBSEs who hold degrees or certifications other than doctorates and suggest possible models for future linkages. The demonstration track aimed to demonstrate the feasibility of acquiring, analyzing, and disseminating linked data files. The two approaches were meant to reinforce each other. This report offers a high-level summary about the NORC project which explored using novel datasets and built a foundation to enable tiered access for data sharing.

Approach and Methods

Research track

The aim of the research track was to gain an understanding of FBSEs. To do this, three novel data sets were identified: the Oak Ridge Institute of Science and Engineering (ORISE) data owned by the Department of Energy (DOE), the Survey of Doctorate Recipients (SDR) owned by the NCSES, and the Association of University Technology Managers Association of Innovation and Marketplace (AIMS) data, owned by the third-party private company, Wellspring.

The first dataset NORC sought to acquire was the ORISE dataset. This dataset contains records of participants and their degrees in federally funded internship and fellowship programs that are part of the Federal STEM Education Strategy.

The second dataset NORC sought to acquire was the SDR dataset. This dataset consists of demographic, education and career history information on individuals with a U.S. research doctorate degree in a science, engineering, or health field. It is a unique source of information about the educational and occupational achievements and career movement of U.S. trained doctoral scientists and engineers in the U.S. and abroad.

Finally, NORC sought to acquire the Wellspring data. This dataset contains records on information on technology transfers between universities and businesses. In addition, the dataset includes records on university-based innovations with names of inventors.

Demonstration track

The demonstration track focused on the acquisition strategies for each dataset with a goal of establishing a data sharing agreement that relied on replicable processes.

The acquisition strategy for the ORISE data included establishing a partnership between NORC and Oak Ridge Associated Universities (ORAU)/ORISE subject matter experts. NORC aimed to gain access to data and data experts under an existing contract between DOE and ORAU/ORISE. However, NORC was unsuccessful in acquiring this dataset within the project period of performance due to legal restrictions which prohibited the sharing of ORISE data for the FBSE project. Some of these concerns involved the limitations associated with having an intermediate organization work to gain access to the DOE-owned ORISE data without the presence of a federal representative. Modifying the existing agreement between DOE with ORAU proved to be complex with multiple parties involved and the period for negotiating the terms of data sharing spanned about a year and a half and was never completed.

The acquisition strategy for the SDR data was straightforward since the federal statistical system, in alignment with Public Law 115-435 §3583, requires statistical agencies to offer a defined and replicable process to request access to confidential data assets. The process for acquiring the SDR data benefited from its standardized processes for data acquisition at NCSES. There were no major barriers encountered to access the SDR data.

The acquisition strategy for the AIMS data was a bit more complex. The strategy included entering into a payable agreement with Wellspring to enable NORC to transfer SDR data for Wellspring to link to the AIMS data and return the appended records (henceforth referred to as covariate data) to NORC. While there were replicable processes that could have been leveraged for the data use agreement and licenses with Wellspring, many of those processes did not address how NCSES would access the linked data when the project concluded since NCSES was not in a data sharing agreement with Wellspring.

To abide by these governance restrictions, NORC developed a tiered access framework strategy that would enable linking the data in two steps. It required that, according to the data sharing agreement between NORC and Wellspring, NORC would provide names from the SDR data to link to the AIMS data. However, both NORC and Wellspring had to become Confidential Information Protection and Statistical Efficiency Act (CIPSEA) designated agents. This strategy presented a unique set of access challenges particularly related to the determination that the confidentiality and privacy of the SDR data were maintained per the CIPSEA statutory requirements. Once the requirements were met, the team began exploring how to link the sources.

To link the dataset, a privacy preserving record linkage (PPRL) tool was explored but not pursued due to limited identifying variables on both sources which would have likely resulted in low quality links. Therefore, an alternative approach was developed to link the sources using clear text linking while incorporating a certain noise infusion technique as a disclosure risk strategy. This linkage resulted in a

crosswalk file with a unique ID that did not include any personally identifiable information between the SDR and AIMS data. The crosswalk file could then be appended to the covariate data from Wellspring to support analyses. However, working through the many layers of the data sharing agreement took much longer than expected and the resultant analytic file with covariate data was unable to be analyzed to inform the research track of this project during the period of performance (note: the final dataset was delivered 22 months after the start of negotiations). Future analyses are possible, but they will require exploring a licensing agreement directly between interested parties and Wellspring for access to the covariate data. This future resource will provide additional information on the FBSE population that is not obtained through SDR.

The results of this process can be viewed as a tiered framework for access to data where the first step is the linkage crosswalk and the second step is obtaining the necessary license/data sharing agreement to add in the proprietary analytic variables but does not rely on the sharing of direct identifiers. This two step/tiered access framework has enabled access to covariate data while removing the need to repeatedly perform linkages to gain access to proprietary data. Ultimately, the work from this project laid the foundation for assessing novel datasets and the data sharing considerations that are needed to link the datasets. Innovative approaches were explored and will lay the groundwork for future data sharing projects.

Considerations for a Potential, Future NSDS

The information gained from this project contributes to the development of an NSDS by offering specific insights into both the legal and technical barriers and processes that facilitate data sharing between entities, including federal and third-party organizations. This project highlighted notable recommendations to help facilitate data access including development of a tiered access framework and the use of privacy enhancing technologies that align with Recommendation 1.7 by the Advisory Committee on Data for Evidence Building Year 2 report. Tiered access frameworks allow for secure sharing of sensitive data and offer solutions for different governance situations.

In summary, this project highlighted limitations, challenges, and opportunities associated with establishing data sharing agreements to support evidence-informed decision making. The implementation of a tiered access framework and the documentation of replicable processes could support the development of an efficient and secure platform for shared services, supporting the possibility of enabling evidence-based research within a future NSDS. The lessons learned will inform future projects that potentially could benefit from data sharing and hopefully limit the time needed to establish the agreements.